

# **Challenges in Cyber Security Experiments: Our Experience**

Annarita Giani, UC Berkeley,  
George Cybenko, Dartmouth College  
Vincent Berk, Dartmouth College  
Eric Renauf , Skaion

# Outline

1. The Situational Awareness system to be tested (Dartmouth)
2. The blind test
3. Ground truth (Skaion)
4. Performance measures (AFRL)
5. Results and conclusion

# Outline

1. The Situational Awareness system to be tested (Dartmouth)
2. The blind test
3. Ground truth (Skaion)
4. Performance measures (AFRL)
5. Results and conclusion

# Process Query System

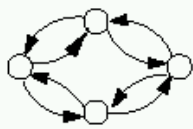
Observable events coming from sensors

Models

Model  $M_1$



Model  $M_2$



...

Model  $M_k$

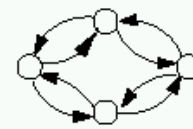


Hypothesis

Likelihood  $L_1$



Likelihood  $L_2$



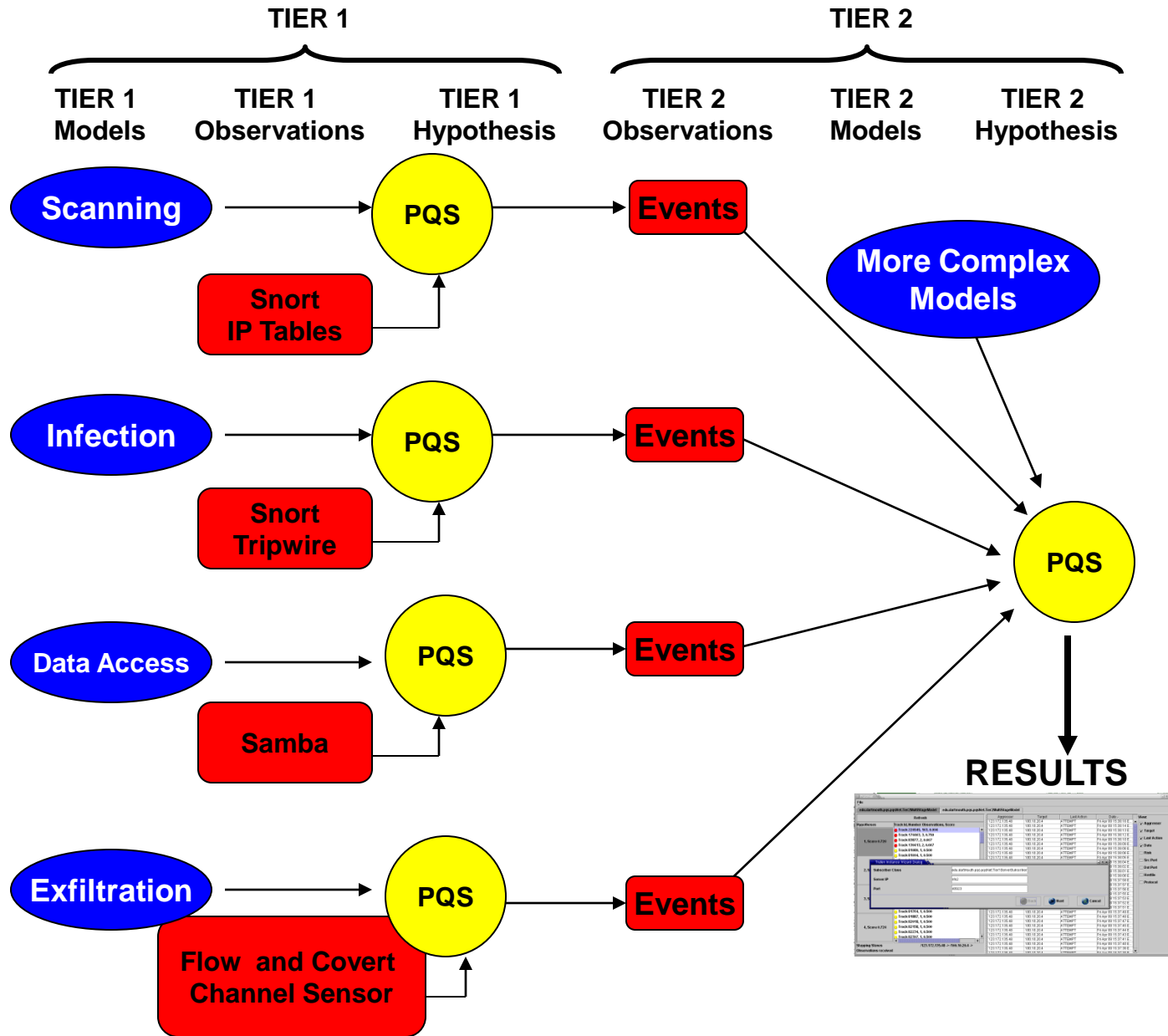
...

Likelihood  $L_k$

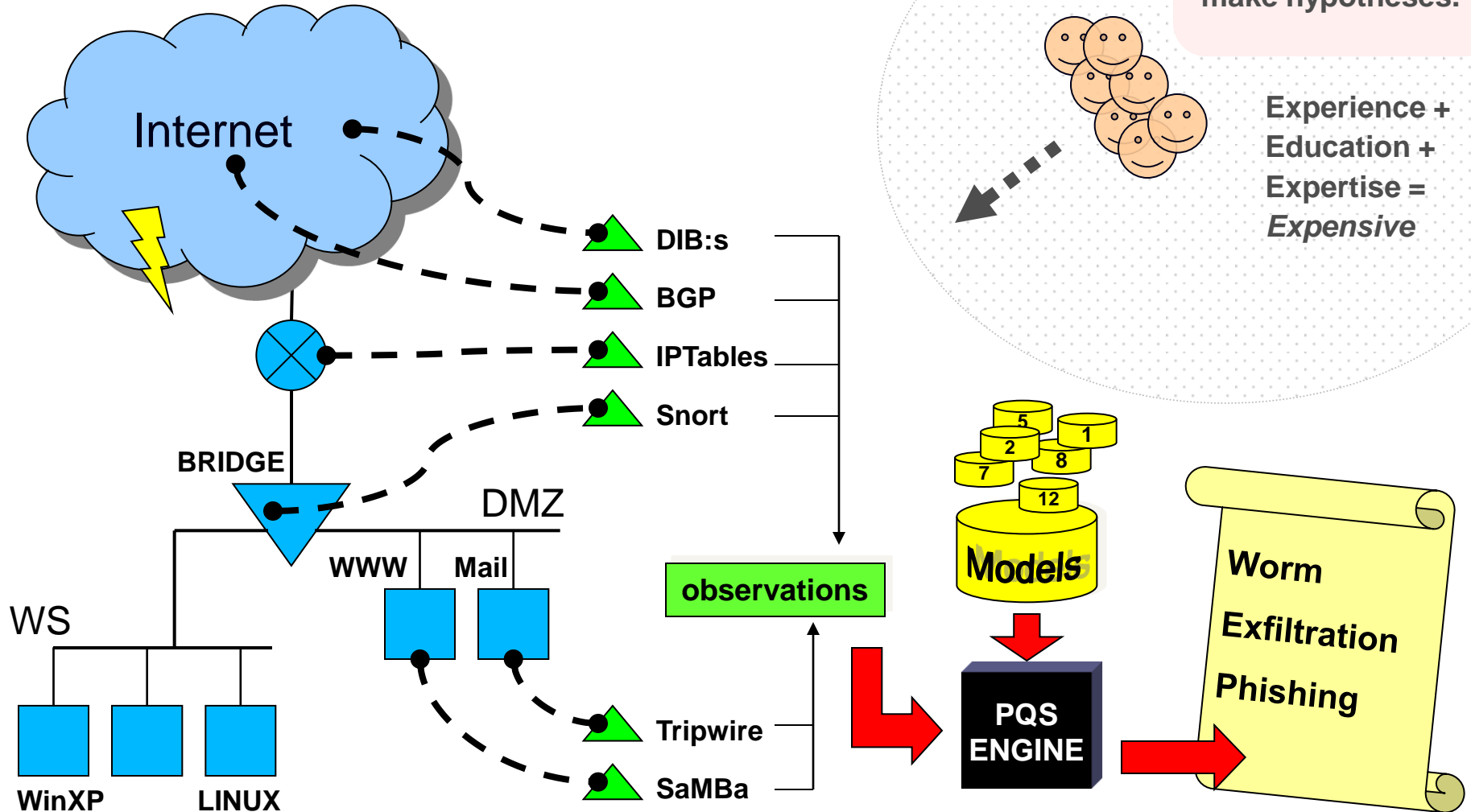


Tracking  
Algorithms








# Hierarchical PQS Architecture










# PQS in Computer Security

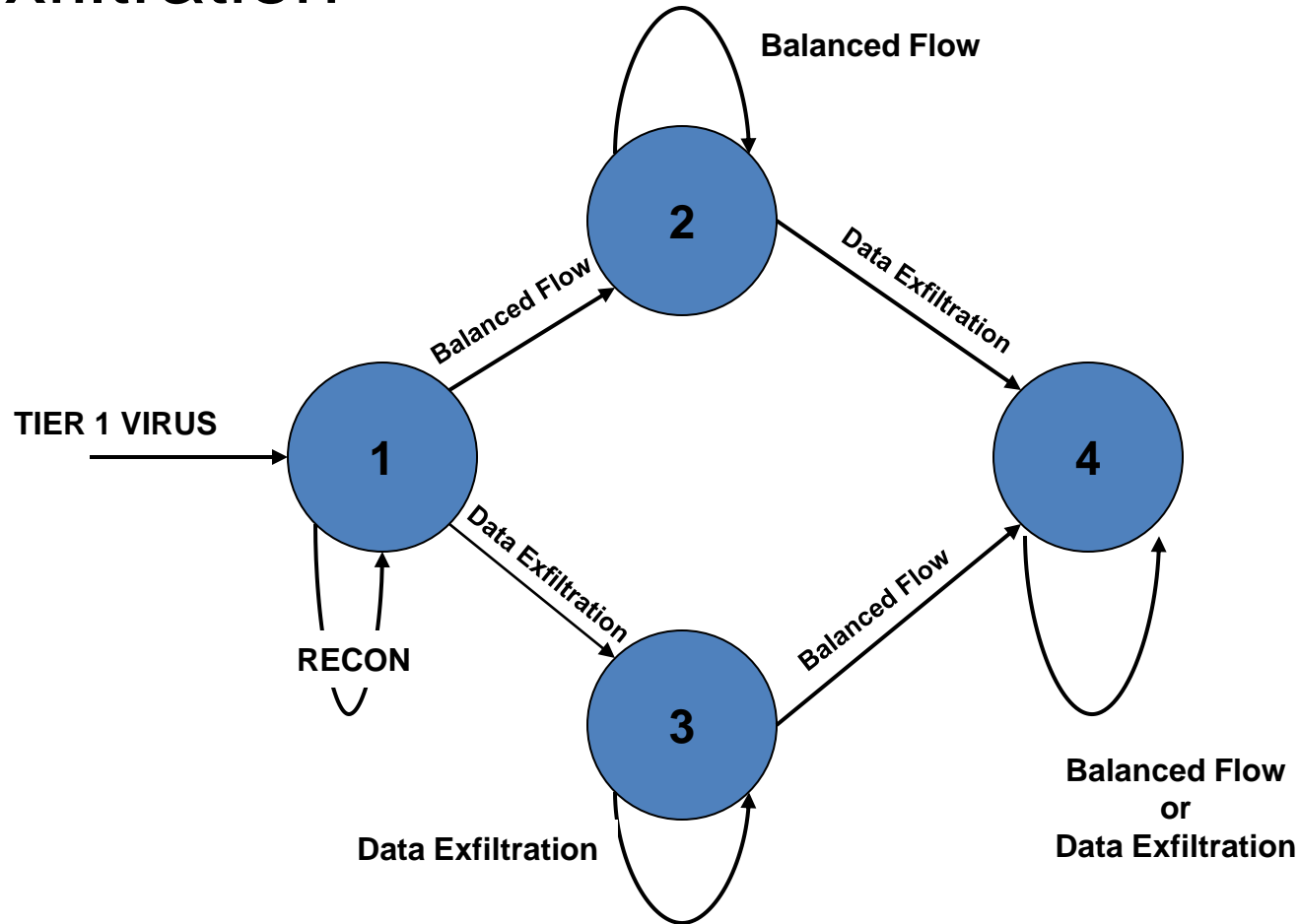


# Sensors and Models

	<b>DIB:s</b>	<b>Dartmouth ICMP-T3 Bcc: System</b>
	<b>Snort, Dragon</b>	<b>Signature Matching IDS</b>
	<b>IPtables</b>	<b>Linux Netfilter firewall, log based</b>
	<b>Samba</b>	<b>SMB server - file access reporting</b>
	<b>Flow sensor</b>	<b>Network analysis</b>
	<b>ClamAV</b>	<b>Virus scanner</b>
	<b>Tripwire</b>	<b>Host filesystem integrity checker</b>

	<b>Noisy Internet Worm Propagation – fast scanning</b>
	<b>Email Virus Propagation – hosts aggressively send emails</b>
	<b>Low&amp;Slow Stealthy Scans – of our entire network</b>
	<b>Unauthorized Insider Document Access – insider information theft</b>
	<b>Multistage Attack – several penetrations, inside our network</b>
	<b>DATA movement</b>
	<b>TIER 2 models</b>

# Example PQS model: Macro in word document for exfiltration



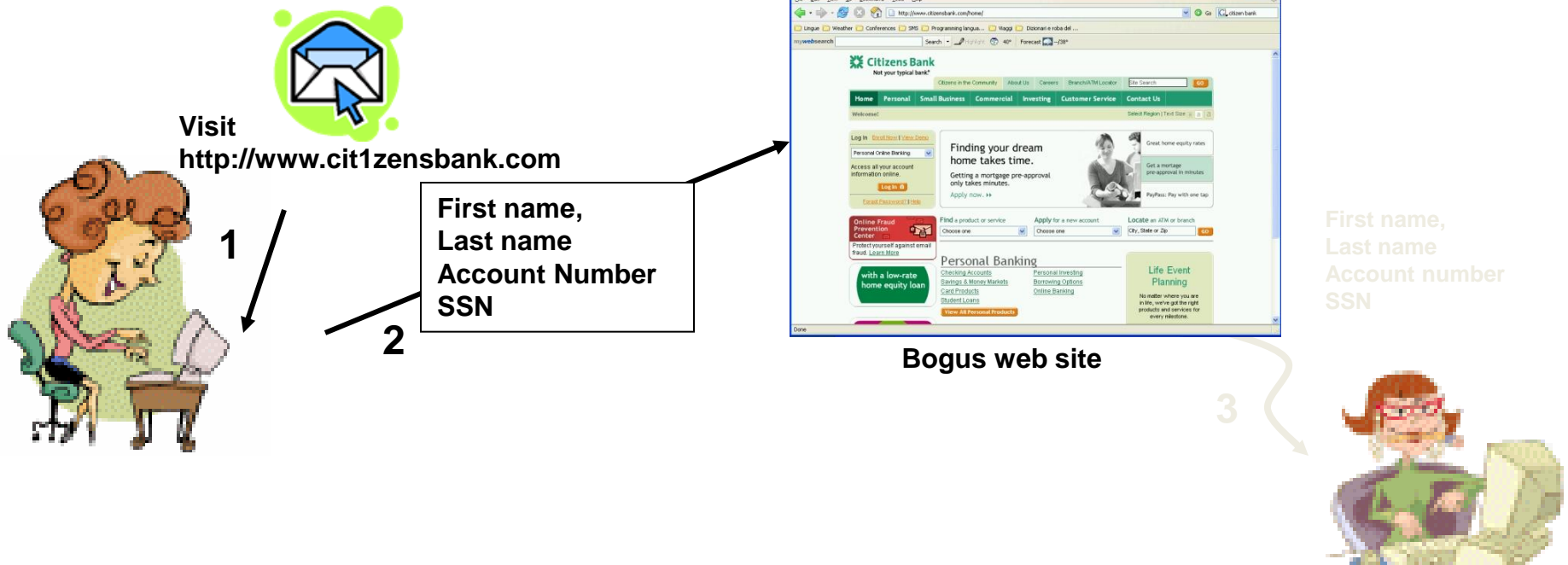
Word virus opens up a ftp connection with a server and upload documents.



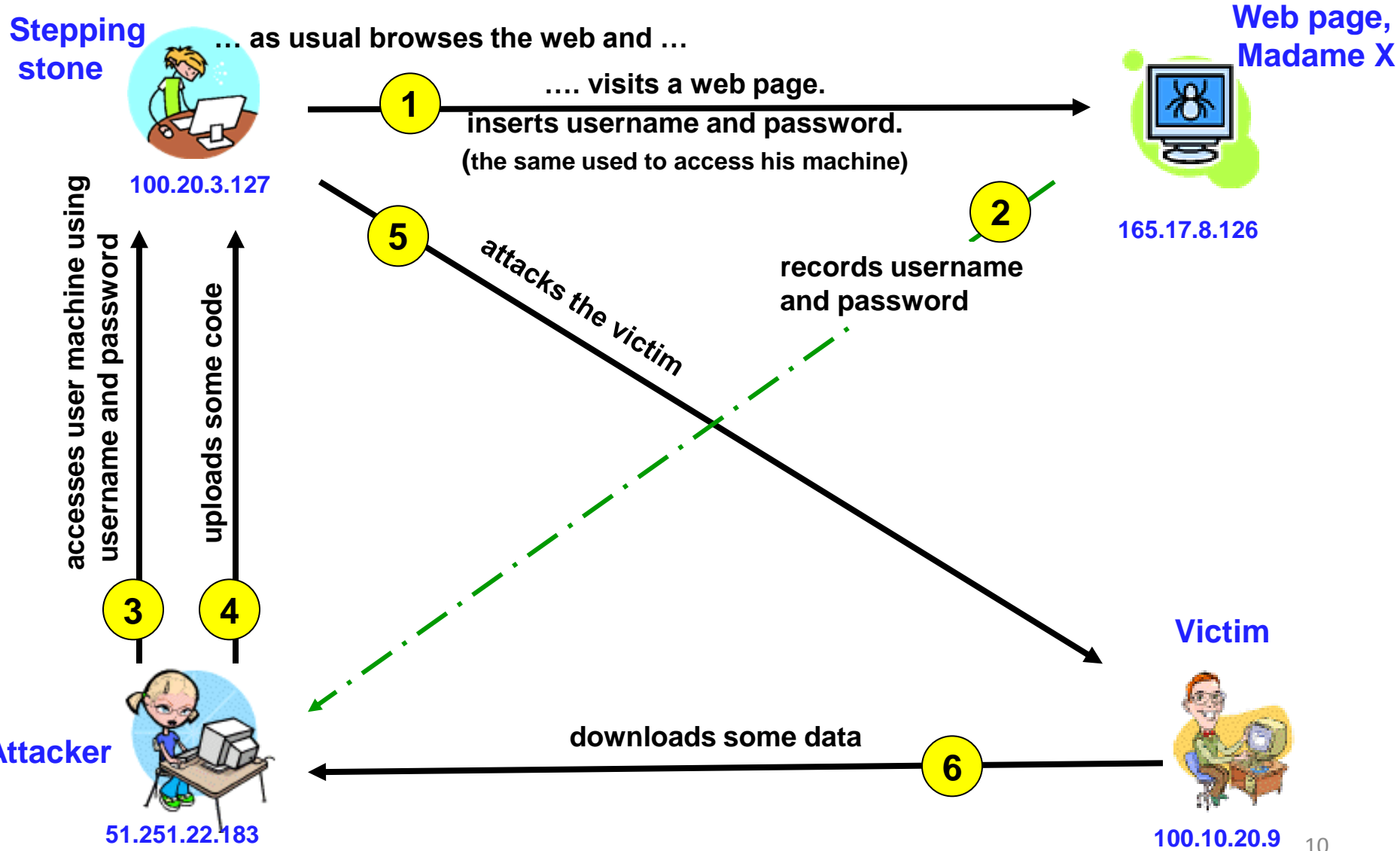
# Phishing Attack

The act of sending an **e-mail** to a user falsely claiming to be an established legitimate **enterprise** in an attempt to scam the user into surrendering private information.

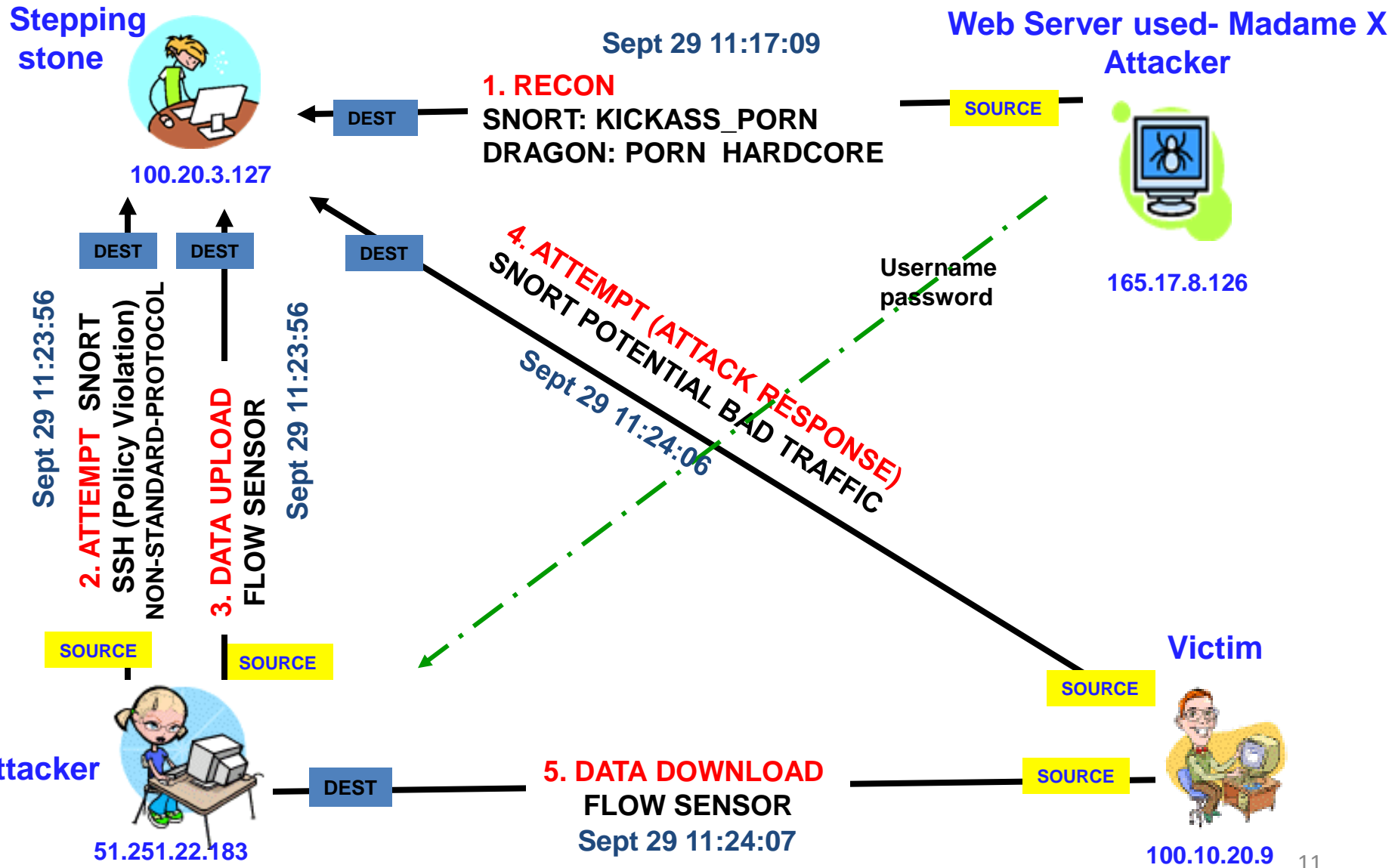
The e-mail directs the user to visit a **web site** where they are asked to update **personal information**.



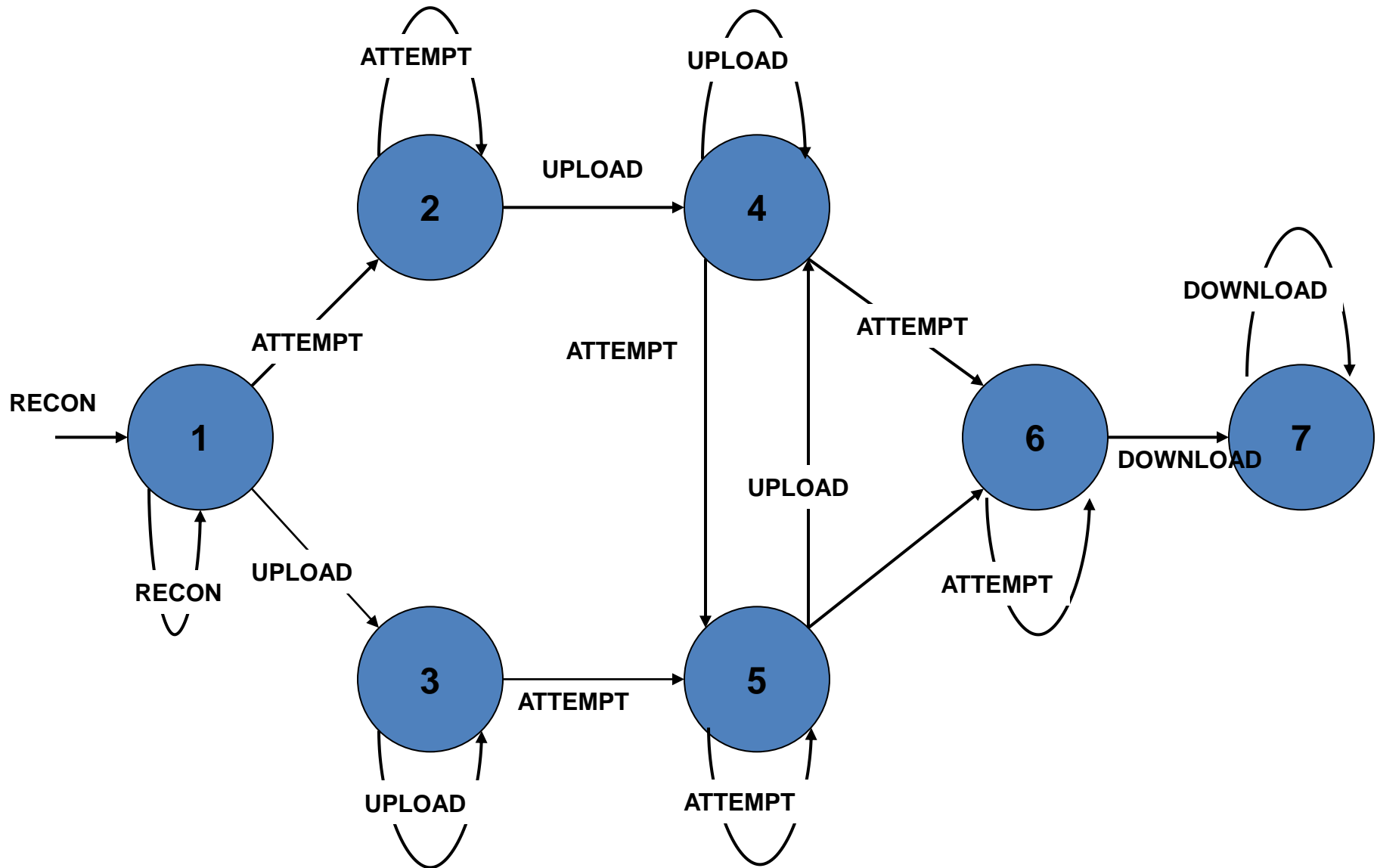
# Complex Phishing Attack Steps



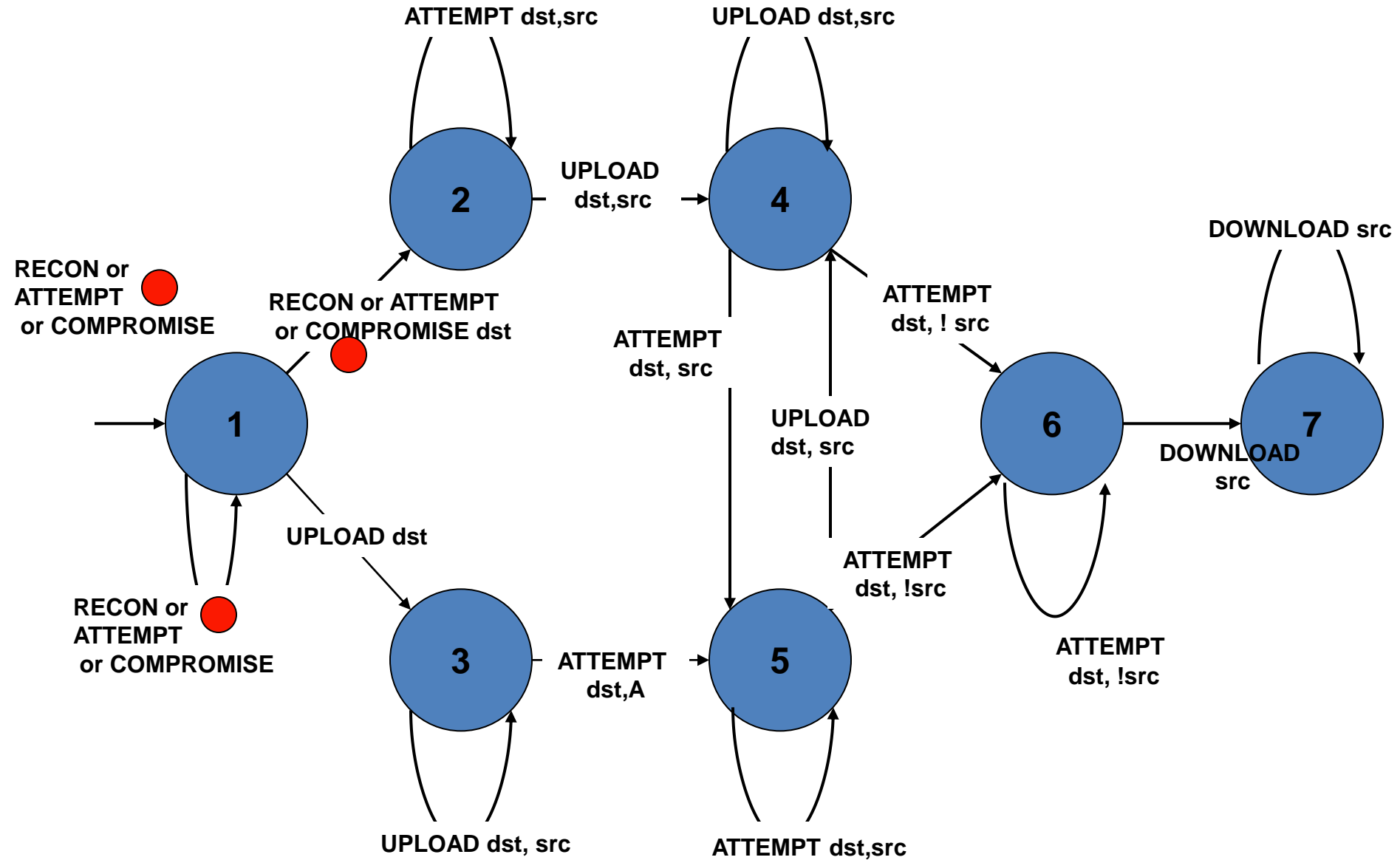
# Complex Phishing Attack Observables



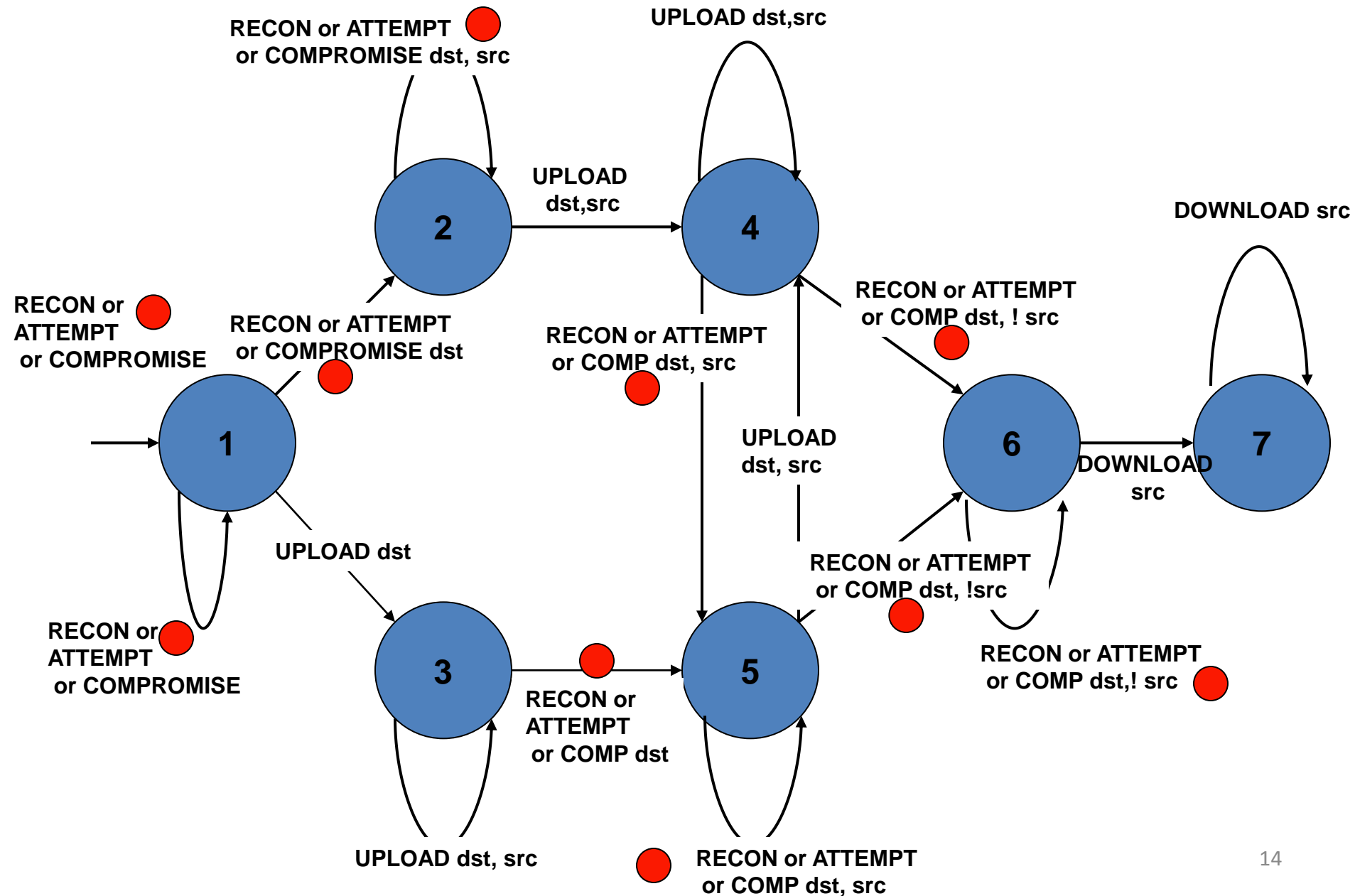
# Phishing Attack Model 1 – very specific



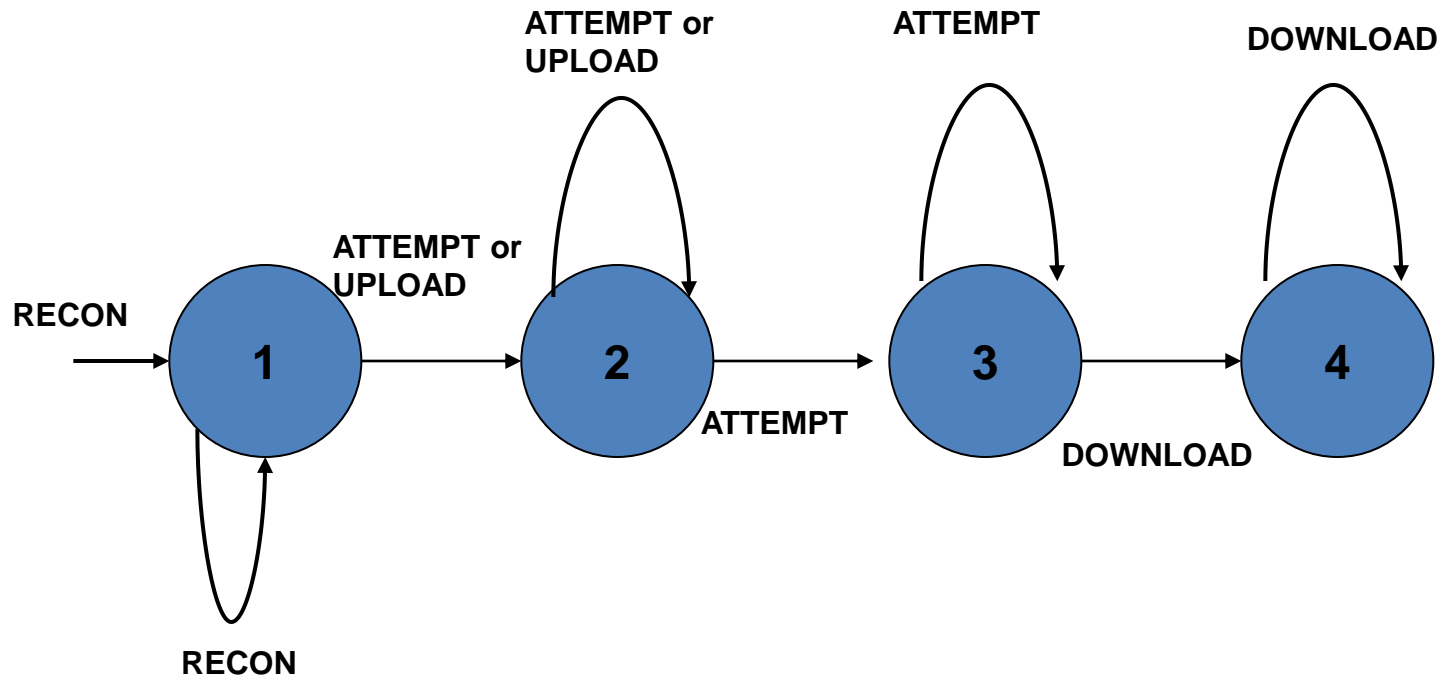
# Phishing Attack Model 2 – less specific



# Phishing Attack Model 3 – more general



# Phishing Attack Model 3 – Most general



Stricter models reduce false positives, **but** less strict models can detect unknown attack sequences

# Outline

1. The Situational Awareness system to be tested (Dartmouth)
- 2. The blind test**
3. Ground truth (Skaion)
4. Performance measures (AFRL)
5. Results and conclusion



# Blind Test

December 12-14, 2005

## Who was there

The system developer team

The reviewers

The ground truth developer team

The SA system developer team were given 4 hard drive full of network and host data. They had to provide the list of the attacks that generated such data. The responses were used to evaluate the systems and guide development.

The collected data is an anonymized stream of network traffic, collected using *tcpdump*. It resulted in hundreds of gigabytes of raw network traffic.

# Outline

1. The Situational Awareness system to be tested (Dartmouth)
2. The blind test
- 3. Ground truth (Skaion)**
4. Performance measures (AFRL)
5. Results and conclusion

# Ground truth (1/2)

**Skaion** provides offline data sets consisting of a variety of captured sensor data streams during simulated scenarios.

During those scenarios, normal background traffic is provided by the Skaion Traffic Generation System (TGS).

An attack as an attempt to gain privileges or information in excess of those granted.

A ground truth system should ideally be able to:

1. handle network-based and host-based alerts.
2. uniquely distinguish every event.
3. categorize all captured behavior, not just malicious behavior.
4. be used by humans or programs to understand the documented scenario.
5. provide multiple flexible perspectives.

# Ground truth (2/2)

Skaion created malicious outsider data sets that included

1. Captured network traffic
2. IDS alerts
3. application log files

Ground truth was represented using **four methods**:

1. A narrative description of attacks
2. a formatted list of attack steps
3. a per-sensor breakdown of false and true positives and negatives
4. a relational database of all captured events

Sam Gordon, Eric Renouf, *Role-Based Ground Truth for Generated Attack Scenarios*  
Skaion Corporation

# Outline

1. The Situational Awareness system to be tested (Dartmouth)
2. The blind test
3. Ground truth (Skaion)
- 4. Performance measures (AFRL)**
5. Results and conclusion

# Definitions

The “raw data” or input streams for a Cyber SA system are the events generated by network sensors.

These events are considered the ***evidence (observations)*** of a Cyber SA system.

When the evidence from multiple data streams is fused together in such a way as to identify a potential attack, we call this collection of evidence ***an attack track***.

*A **situation** is the set of all attack tracks at a given point in time.*

From Tadda: *Measuring Performance of Cyber Situation Awareness Systems*, 11th International Conference on Information Fusion 2008

John Salerno, *Measuring situation assessment performance through the activities of interest score*, 11th International Conference on Information Fusion 2008

These definitions were unknown at the test.

# Metrics of Performance

The metrics fall into four dimensions:

1. confidence
2. purity
3. cost utility
4. timeliness

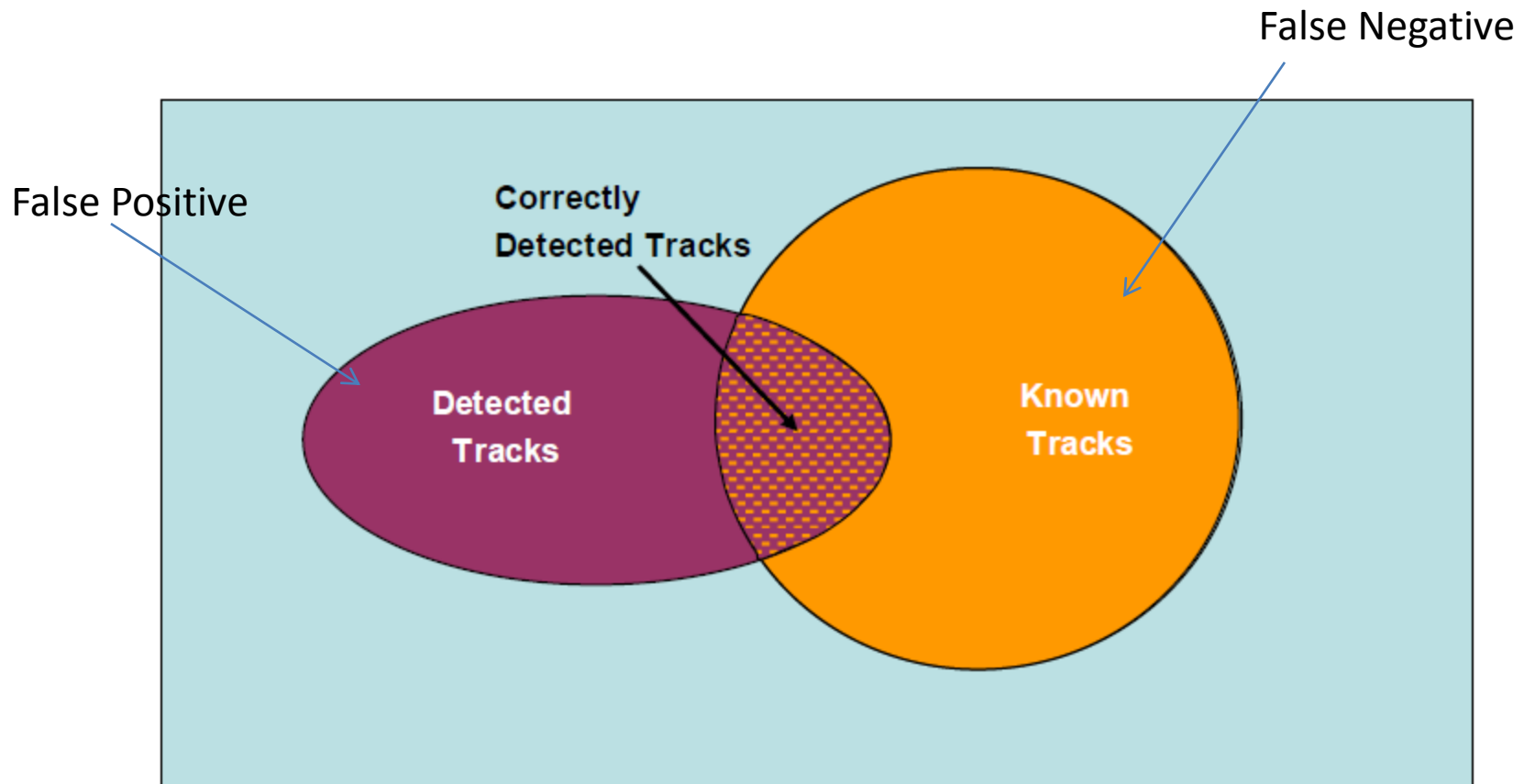
[Tadda: Measuring performance of cyber situation awareness systems](#) 11th

International Conference on Information Fusion, 2008

# Confidence

How well the system detects the true attack tracks.

- (1) recall
- (2) precision
- (3) fragmentation
- (4) mis-association





# Purity

Quality of the correctly detected tracks. How well the evidence is being correlated and aggregated into an attack track.

$$\text{MISASSAGNEMENT RATE} = \frac{\text{Number Of Incorrect Evidence}}{\text{Total Evidence Detected}}$$

whether the system was assigning evidence to a track that wasn't relevant or if it only considered directly useful evidence

$$\text{EVIDENCE RECALL} = \frac{\text{Number Of Correct Evidence}}{\text{Total Evidence in Ground Truth}}$$

How much of the observation available was truly being used?

# Cost Utility

the ability of a system to identify the “important or key” attack tracks with respect to the concept of cost. In [8], two cost utility metrics were described.

$$\text{ATTACK SCORE} = \frac{\text{NAGT} \times \text{NTGT} - \sum_{i=1}^{\text{NAD}} P_i}{\text{NAGT} \times \text{NTGT} - \sum_{i=1}^{\text{NTGT}} i}$$

NAGT    Number of Attacks in the Ground Truth

NTGT    Number of Tracks in the Ground Truth

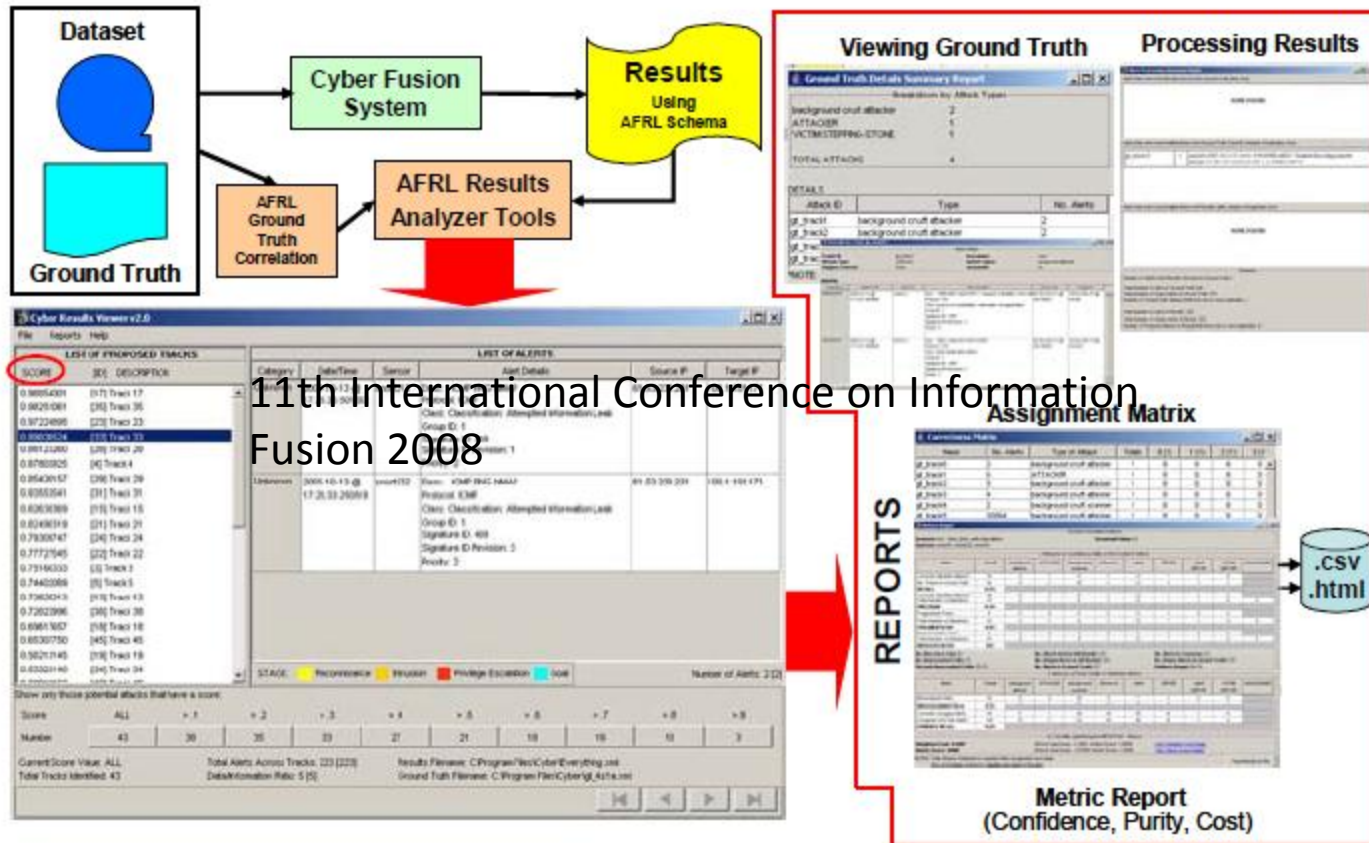
NAD     Number of Attacks Detected

$P_i$      position of the  $i^{th}$  attack in the results

# Timeliness

The ability of the system to respond within the time requirements of a particular domain.

# Assessment Process



11th International Conference on Information Fusion 2008

# Outline

1. The Situational Awareness system to be tested (Dartmouth)
2. The blind test
3. Ground truth (Skaion)
4. Performance measures (AFRL)
- 5. Results and conclusion**

# Complex Phishing Attack Results

No observations coming from Dragon sensor and Flow sensor

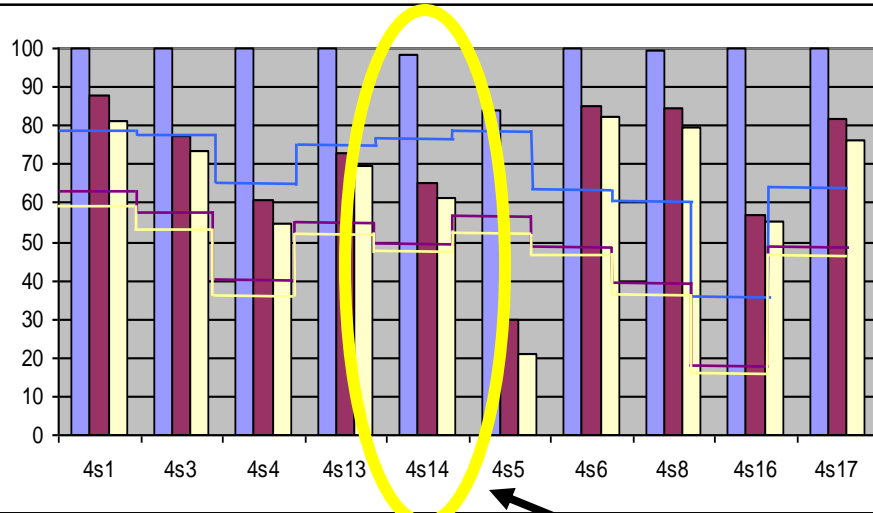
Attack steps	0 of 5
Background attackers	9 of 15
Background scanners	25 of 55
Stepping stones	0 of 1

Using Dragon and Flow observations

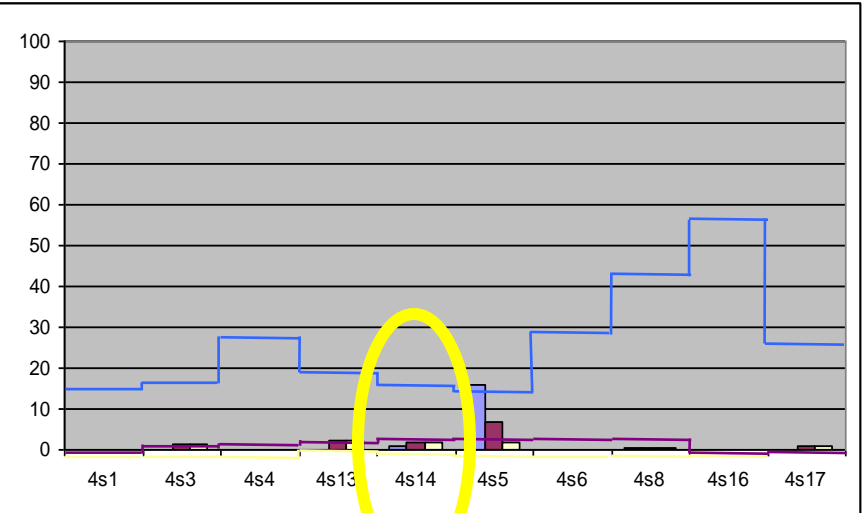
Attack steps	5 of 5
Background attackers	10 of 15
Background scanners	23 of 55
Stepping stones	1 of 1
False alarms	1

# Summary of Results

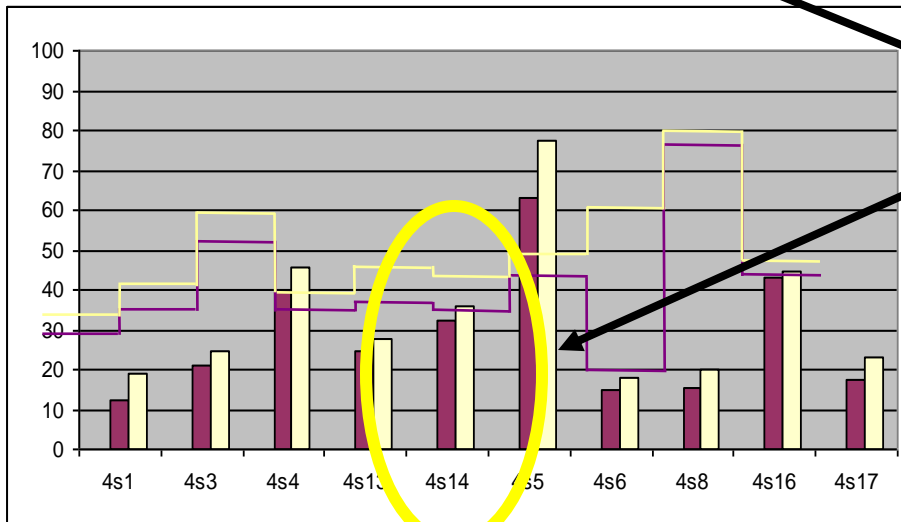
## Precision



## Fragmentation



## Mis-Associations



Scenario 4s14: Phishing attack

Threshold Values: 0.0 0.5 0.75

# Putting Everything Together

Valuable feedback on performance and design

(Animated) Disagreement on the definition of what an attack is.  
This can lead to lead to different degrees of performance.

Sleepless nights trying to have the results in the format the reviewer wanted so that they could run their software assessment.

Are the precision measurements proposed by the reviewers good? How do we assess that? This lead to the problem of choosing right measures of performance.

